A DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS GEOGRÁFICAS ATRAVÉS DA EXPLICITAÇÃO SEMÂNTICA

MARIBEL SANTOS **LUÍS AMARAL** **PEDRO PIMENTA**

NÚCLEO DO DEPARTAMENTO DE INFORMÁTICA
UNIVERSIDADE DO MINHO
CAMPUS DE AZURÉM
4800 GUIMARÃES
PORTUGAL
TELF.: +351 53 510259
FAX: +351 53 510250
{maribel, amaral, pimenta}@dsi.uminho.pt

RESUMO

A investigação na área da Descoberta de Conhecimento em Bases de Dados Geográficas tem sido caracterizada pelo desenvolvimento de algoritmos de *Data Mining* capazes de captar e utilizar a semântica associada à componente espacial dos dados analisados. Este artigo apresenta uma nova abordagem na qual é possível a utilização de algoritmos de *Data Mining* já disponíveis no mercado e não desenvolvidos para lidar com dados geográficos. Esta aproximação baseia-se na explicitação semântica de alguns dos relacionamentos espaciais existentes entre as entidades analisadas, como seja a direcção ou distância entre elas. A explicitação é conseguida utilizando os princípios definidos nas normas CEN TC 287 para Informação Geográfica. A partir das suas directivas e baseado num conjunto reduzido de relacionamentos é possível a inferência de novos relacionamentos espaciais desconhecidos para o sistema. A abordagem proposta foi validada partindo de uma base de dados geográfica contendo a orientação espacial existente entre alguns Concelhos de Portugal, a qual permitiu a inferência da direcção actual entre os Distritos que agregam os Concelhos analisados.

ABSTRACT

Knowledge Discovery in Geographic Databases has been characterised by the development of new algorithms able to catch and use the semantic associated with the fact's locations. This paper presents a new approach in that is possible the use of Data Mining algorithms not implemented for geographic data treatment and already available in market. It is based in the semantic explicitation of some spatial relations that exist between the analysed entities, as the direction or distance among them. This explicitation is reached using the principles established by the CEN TC 287 Geographic Information standard. Under its directives and based on a small set of relations is possible the inference of new spatial relations. The proposed approach was validated using a geographic database with the spatial directions that exist between some Municipalities of Portugal, allowing the inference of the direction relation among the Districts that aggregate the analysed Municipalities.

1 INTRODUÇÃO

O volume de dados armazenados e manipulados pela maioria das organizações cresce diariamente a uma taxa que ultrapassa a nossa capacidade de analisar, sintetizar e extrair conhecimento a partir desses dados. Apesar dos Sistemas Gestores de Bases de Dados (SGBD) fornecerem ferramentas capazes de armazenar e visualizar grandes quantidades de dados, a utilização de ferramentas específicas, desenhadas e implementadas com o objectivo de automatizar o processo de análise dos dados, é cada vez mais imprescindível.

É neste contexto que surge uma nova área de investigação denominada *Knowledge Discovery in Databases*. O processo de Descoberta de Conhecimento em Bases de Dados (DCBD), que se desenvolve em várias fases, inclui a gestão dos algoritmos de *Data Mining*, utilizados para extrair padrões dos dados, e a interpretação dos padrões encontrados pelos mesmos. As ferramentas de DCBD utilizam uma diversidade de algoritmos para identificar relacionamentos e padrões que estão escondidos entre o grande volume de dados. Os relacionamentos e padrões encontrados representam conhecimento acerca da Base de Dados (BD) explorada e das entidades nela contida. Decidir

se os achados reflectem ou não conhecimento útil, é uma das fases do processo na qual a participação do utilizador é normalmente necessária [Fayyad, et al. 1996a].

Os grandes progressos conseguidos até ao momento na área da DCBD têm-se restringido quase que exclusivamente à exploração de dados armazenados em BD relacionais. Existe contudo na maioria das BD organizacionais uma dimensão dos dados, a geográfica (explícita normalmente através de uma morada, código postal, etc.), cuja semântica não é utilizada por estas técnicas de DCBD.

As técnicas de Descoberta de Conhecimento em Bases de Dados Geográficas (DCBDG) desempenham um papel fundamental na percepção dos dados geográficos e na identificação dos relacionamentos implícitos existentes entre os dados geográficos e os dados não geográficos.

A semântica associada à localização dos factos, a análise dessas localizações com o objectivo de perceber o seu porquê, faz com que a utilidade do conhecimento obtido através do processo de descoberta de conhecimento seja largamente melhorada com a integração de dados geográficos e dados não geográficos.

Este trabalho visa a incorporação e utilização da dita semântica no processo de DCBD. O objectivo é utilizar ferramentas de DCBD já disponíveis no mercado. Toda a semântica será especificada utilizando as normas para Informação Geográfica (IG) desenvolvidas pelo Comité Europeu de Normalização (CEN). As normas CEN TC 287 desempenham um papel fundamental na definição e estruturação da informação utilizada neste estudo. Têm como principal objectivo permitir que a IG possa ser acedida por diferentes utilizadores, sistemas, aplicações e principalmente, de diferentes localizações. Para tal é necessário definir e descrever IG de uma forma padronizada, definir os métodos e estruturas para o seu armazenamento e ainda, definir como esta informação pode ser acedida, pesquisada e transferida [CEN/TC-287 1996b].

Salienta-se que a referenciação geográfica da informação utilizada neste estudo é conseguida recorrendo a um **sistema de identificadores geográficos**. Este sistema é especificado segundo as recomendações da norma e tem como grande vantagem o facto de permitir a inferência de informação espacial implícita, a partir de um pequeno conjunto de relacionamentos espaciais explícitos no sistema. Os relacionamentos espaciais considerados nesta primeira fase do estudo dizem respeito a **direcção** existente entre entidades, a qual é expressa através de um identificador qualitativo como *Norte*, *Sul*, …

2 O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A década de noventa trouxe consigo um notável crescimento da quantidade de informação gerada e armazenada pela maioria das organizações. A necessidade de recolher e armazenar dados de diversos tipos e proveniências superou a nossa capacidade de analisar, sintetizar e extrair conhecimento a partir desses dados. Para lidar com esta emergência de informação são necessárias ferramentas inteligentes que automatizem o processo de análise dos dados e descoberta de conhecimento.

A DCBD (*Knowledge Discovery in Databases*) é definida como *"o processo não trivial de identificação de padrões a partir dos dados, sendo os mesmos válidos, desconhecidos e potencialmente úteis"* ([Fayyad, et al. 1996b] pág.6). Os algoritmos utilizados para extrair padrões dos dados são denominados de *Data Mining*. O processo global de DCBD (Figura 1), que se desenvolve em várias fases, inclui a gestão dos algoritmos de *Data Mining* e a interpretação dos padrões encontrados pelos mesmos, os quais serão utilizados posteriormente no suporte à tomada de decisão.

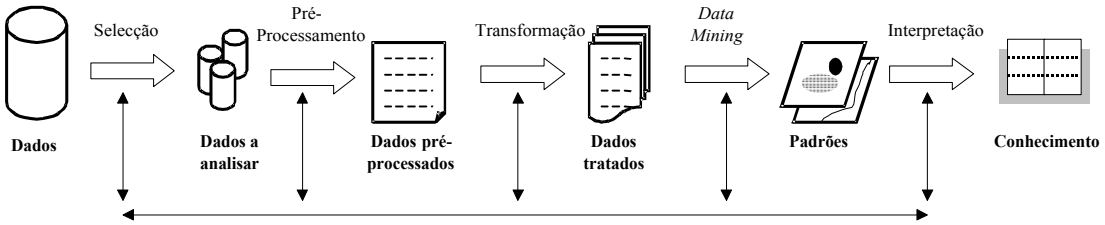

Figura 1 – O processo de Descoberta de Conhecimento em Bases de Dados (Adaptado de: Fayyad, U., G. Piatetsky-Shapiro, e P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, 39, 11 (1996), 27-34, p. 29)

Um dos principais problemas com que se deparam as técnicas de *Data Mining* é que o número de possíveis relacionamentos é extremamente elevado, ocultando por vezes os mais importantes. As estratégias de pesquisa têm então de ser inteligentes, para o qual se recorre a área da Aprendizagem Automática (*Machine Learning*). Outro problema encontrado com bastante frequência é a existência de dados corrompidos ou desconhecidos, os quais conduzem normalmente à utilização de técnicas estatísticas para avaliar o grau de confiança dos relacionamentos encontrados [Holsheimer e Siebes 1994].

Basicamente, as diversas etapas que caracterizam o processo de DCBD podem ser sintetizadas em:

1. **Aprendizagem** do domínio de aplicação, que inclui a percepção do conhecimento relevante sobre o domínio e os objectivos a atingir no processo;
2. **Selecção** dos dados a tratar, sobre os quais incidirão os algoritmos de *Data Mining*;
3. **Pré-processamento** para verificação dos dados. É frequente o aparecimento de dados incorrectos ou valores omissos, para os quais é necessário definir uma estratégia de actuação. Pode, ainda, ser constatada a existência de atributos redundantes ou a falta de atributos relevantes que conduzem necessidade de integração de novos dados;
4. **Transformação** dos dados, o que inclui a procura de configurações apropriadas para representar os dados. A utilização de mecanismos de transformação dos dados tem como objectivo diminuir o número de registos e/ou atributos em análise;
5. Verificar quais as **técnicas de *Data Mining*** disponíveis (classificação, regressão, *clustering*, etc.) e que satisfazem os objectivos pretendidos;
6. Escolher o(s) **algoritmo**(s) de *Data Mining* a utilizar;
7. ***Data Mining***, para a procura dos padrões;
8. **Interpretação** dos padrões encontrados. Nesta fase inclui-se a visualização dos padrões encontrados, a remoção de padrões irrelevantes ou redundantes, ou simplesmente a transformação dos desejados para formatos perceptíveis ao utilizador (os resultados encontrados por alguns algoritmos, como por exemplo as redes neuronais, têm de ser tratados por forma a poderem ser efectivamente utilizados).

Todo este processo é iterativo e interactivo possibilitando retrocessos nas diversas fases e a participação do utilizador sempre que se justifique a tomada de decisão. Em termos de "trabalho", a fase de *Data Mining* representa normalmente 20% do tempo gasto em todo o processo. Esta é também a fase que é melhor suportada automaticamente (por *software*). Todas as outras fases, desde a selecção dos dados até a interpretação dos padrões encontrados, constituem mais uma questão de "arte" do que uma rotina que possa ser automatizada [Andrienko e Andrienko 1998].

3 APROXIMAÇÕES À DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS GEOGRÁFICAS

Na área dos dados espaciais as principais aproximações à descoberta de conhecimento incluem a utilização de técnicas de generalização para derivação de características existentes entre os dados [Lu, et al. 1993]. Esta derivação

pode ser iniciada a partir de dados geográficos, os quais são por exemplo agregados até se atingir o nível de generalização desejado. A partir das classes encontradas, os dados não geográficos são analisados e as principais características dos mesmos identificadas com o objectivo de serem associadas as classes geográficas encontradas. Uma outra abordagem possível é partir dos dados não geográficos, os quais são agregados por forma a encontrar classes, às quais são posteriormente associadas características geográficas.

Salienta-se que a aproximação anterior é conseguida através do desenvolvimento de *software* específico que permita a análise geográfica da informação analisada, ou então, do desenvolvimento de uma camada aplicacional que permita a integração do sistema de descoberta de conhecimento com um Sistema de Informação Geográfica (SIG), por forma a que dados geográficos e dados não geográficos possam ser integrados no processo.

Koperski et al. [Koperski e Han 1995; Koperski e Han 1996; Koperski, et al. 1998] investigam a utilização de técnicas de *Data Mining* interactivas (com pesquisas à BD em GMQL – *GeoMining Query Language*) que permitem, entre outras, a descoberta de regras de associação geográfica. Nesta abordagem dados geográficos e não geográficos são armazenados em diferentes BD, mas a partir do momento em que o utilizador especifica o critério de pesquisa, inicia-se um processo de selecção de dados, onde duas camadas de *software* se complementam na criação de um repositório único que contenha os dados a tratar.

Dey e Roberts [Dey e Roberts 1996] adoptam uma aproximação que estende o modelo genérico de DCBD proposto por Matheus et al. [Matheus, et al. 1993], através da utilização de duas interfaces, uma para aceder a dados geográficos e outra para aceder a dados não geográficos (armazenados em diferentes BD). A integração das diferentes BD é lógica, e é conseguida pelo próprio processo de DCBD. É assumido que a BD não geográfica contém identificadores geográficos, como por exemplo moradas, que especificam localizações na BD geográfica. O processo de descoberta de conhecimento pode ser iniciado a partir de qualquer uma das BD.

As aproximações à descoberta de conhecimento em dados geográficos referidas anteriormente, ainda que representem progressos significativos na área, apresentam limitações ao nível da análise produzida. Isto porque os resultados obtidos baseiam-se apenas na procura de padrões naqueles atributos especificados a partida pelo utilizador. Em todas as abordagens referidas, o utilizador é confrontado com a necessidade de especificar um critério de pesquisa, o qual identifica os atributos a analisar e ainda o tipo de resultados pretendido. O facto de ter de existir a partida, uma especificação daquilo que se pretende, limita o leque de resultados que podemos encontrar. Salienta-se que existem neste momento diversas ferramentas de DCBD, onde não é necessária a explicitação de uma questão que "guie" o processo de descoberta de conhecimento. Todos os atributos resultantes das fases de selecção, pré-processamento e transformação dos dados encontram-se armazenados numa única tabela, sobre a qual incidem os algoritmos de *Data Mining*. Esta estratégia tem como grande vantagem o facto da procura não ser direccionada, o que permite que se encontrem padrões, onde o utilizador menos espera.

A abordagem proposta neste trabalho não inclui o desenvolvimento de novos algoritmos adaptados à componente geográfica dos dados [Ester, et al. 1998a; Ester, et al. 1997; Ester, et al. 1998b; Ester, et al. 1995; Koperski e Han 1995; Koperski, et al. 1998; Lu, et al. 1993], ou a implementação de uma nova camada de *software* responsável pela integração de dados geográficos e dados não geográficos [Abraham e Roddick 1997; Abraham e Roddick 1998; Dey e Roberts 1996], mas passa pelo aproveitamento das capacidades de análise exploratória dos dados conseguidas até ao momento pelas ferramentas já desenvolvidas para BD relacionais. Esta aproximação recorre à explicitação semântica, numa BD relacional, de alguns dos relacionamentos espaciais existentes entre os identificadores geográficos utilizados para localizar uma dada entidade geograficamente. A partir deste conjunto reduzido de informação, é possível inferir informação espacial sem a necessidade de recorrer a um SIG ou ao desenvolvimento de novos algoritmos. Esta abordagem é apresentada com mais detalhe na secção 5.

4 AS NORMAS CEN TC 287 PARA INFORMAÇÃO GEOGRÁFICA

As normas CEN TC 287 para IG desempenham um papel fundamental na definição e estruturação da informação utilizada neste estudo. Têm como principal objectivo permitir que a IG possa ser acedida por diferentes utilizadores, sistemas, aplicações e principalmente, de diferentes localizações. Para tal é necessário definir e descrever IG de uma forma padronizada, definir os métodos e estruturas para o seu armazenamento e ainda, definir como esta informação pode ser acedida, pesquisada e transferida [CEN/TC-287 1996b] (ressalva-se que neste momento a norma encontra-se em fase de votação final pelo CEN, pelo que na realidade é apelidada de pré-norma, apesar de ao longo de todo o artigo ser sempre referida como norma).

A adopção desta norma, cuja utilização neste trabalho é estratégica visto Portugal ser membro do referido comité e como tal ter-se comprometido a adoptar as suas resoluções, terá ainda a vantagem de permitir:

- Clarificar os conceitos associados à utilização e integração de IG;
- Aumentar a disponibilidade de IG, incluindo informação acerca da informação geográfica (meta-informação);
- Permitir a transferência de IG e consequentemente a sua reutilização para diferentes propósitos.

A descrição dos dados, seguindo esta norma, tem como principal objectivo permitir a comunicação de vários intervenientes dentro de um sistema de informação ou entre sistemas de informação. Esta descrição especifica a estrutura dos dados e clarifica a sua semântica através da **modelação**. Este processo inclui a construção de um **esquema de aplicação** (resultante da integração de esquemas específicos definidos pela norma, como pode ser constatado pela análise da Figura 2) que define e descreve os dados, assim como regras aplicáveis aos mesmos.

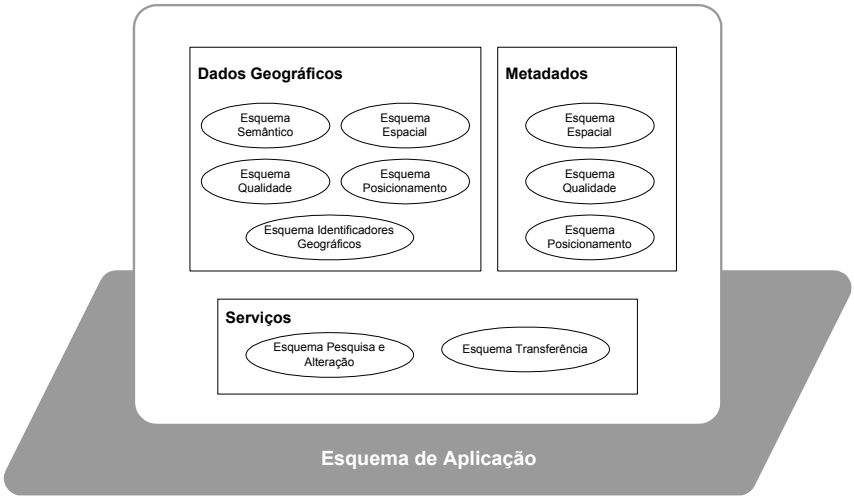

Figura 2 – Esquemas conceptuais que integram um dado Esquema de Aplicação

A linguagem escolhida pelas normas para a definição dos esquemas é o EXPRESS (ISO 10303-11) [Schenck e Wilson 1994]. Esta linguagem permite modelar o esquema de aplicação, independentemente de qualquer implementação física. Os principais componentes utilizados na modelação são: entidades, atributos, relacionamentos, regras, tipos de dados, funções e procedimentos [CEN/TC-287 1998e]. Salienta-se ainda que a definição dos requisitos a implementar para cada esquema também se encontram especificados em EXPRESS.

As normas identificam duas categorias de dados: dados geográficos e metadados. Dados geográficos são representações computadorizadas de IG que incluem **aspectos semânticos** (descrição do modelo a representar),

aspectos espaciais (descrição do posicionamento e forma) e **aspectos de qualidade** (descrição do potencial dos dados) [CEN/TC-287 1996b].

Os **aspectos semânticos** incluem a identificação das entidades, dos atributos que caracterizam cada entidade e dos relacionamentos existentes entre as mesmas. A esquematização de toda esta informação, para um dado domínio de aplicação, dá origem ao esquema semântico [CEN/TC-287 1996b]. Este esquema contém o modelo conceptual através do qual diversos utilizadores podem adquirir um entendimento comum acerca da aplicação em causa.

Os **aspectos espaciais** permitem que dados geográficos:

- De diferentes origens possam ser integrados através da geo-referenciação;
- Sejam analisados utilizando operadores e funções espaciais, permitindo a pesquisa de informação a partir da informação já conhecida;
- Sejam representados graficamente em mapas.

Estes aspectos incluem a utilização de estruturas especializadas, atributos e relacionamentos para o posicionamento e para características geométricas e topológicas. O sistema de posicionamento é representado por um modelo matemático que permite atribuir coordenadas a uma dada localização à superfície terrestre. Um sistema de posicionamento indirecto permite identificar uma localização através da utilização de uma morada ou outro identificador geográfico. Os aspectos espaciais dão origem ao esquema espacial [CEN/TC-287 1996a], esquema de posicionamento [CEN/TC-287 1998g] e ao esquema de identificadores geográficos [CEN/TC-287 1998h].

Os **aspectos de qualidade** permitem determinar a usabilidade de um conjunto de dados e previnem a utilização incorrecta dos mesmos (os resultados da análise espacial dependem da qualidade dos dados utilizados). Os aspectos de qualidade complementam os aspectos semânticos e espaciais descritos anteriormente. Dão origem ao esquema de qualidade, o qual se encontra dividido em três grandes categorias: precisão (*accuracy*), exactidão (*completeness*) e actualidade (*up-to-dateness*). O documento prENV 12656 [CEN/TC-287 1998b] limita-se a definir um modelo através do qual o utilizador pode avaliar a qualidade da informação que lhe é fornecida, confrontando-a com a sua especificação. Tal procedimento permitirá ao utilizador verificar se os dados fornecidos satisfazem realmente os seus objectivos.

A descrição dos metadados (através do esquema de metadados) permite a especificação das características dos dados em questão, como sejam a identificação do proprietário, do conteúdo e da estrutura dos dados, e ainda a disponibilidade dos mesmos. O esquema de metadados é construído recorrendo a três esquemas: espacial, de qualidade e de posicionamento. O documento prENV 12657 [CEN/TC-287 1998a] define o esquema conceptual para especificação dos metadados de dados geográficos. Salienta-se que o documento define qual o conjunto mínimo de metadados que são necessários para descrever os dados, não se preocupando com a implementação física da BD que deverá armazenar os metadados.

Além dos aspectos referidos anteriormente, e que permitem descrever os dados e os metadados, o comité especificou ainda, os princípios que devem reger os **serviços de pesquisa, actualização e transferência** de IG entre sistemas.

Os serviços de transferência dão origem ao esquema de transferência, através do qual dados e metadados podem ser transferidos entre diferentes sistemas. Os esquemas de transferência têm como principal objectivo definir os moldes em que a troca de dados pode ser efectuada [CEN/TC-287 1998c].

O serviço de pesquisa e actualização permite a um utilizador colocar questões e receber respostas relacionadas com os dados geográficos e metadados, utilizando interacção directa ou indirecta através dos serviços de transferência. A especificação deste serviço dá origem ao esquema de pesquisa e actualização [CEN/TC-287 1998f].

5 A EXPLICITAÇÃO SEMÂNTICA NO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS GEOGRÁFICAS

As normas CEN TC 287 requerem a implementação de oito esquemas, que são integrados posteriormente, dando origem ao esquema de aplicação [CEN/TC-287 1998d], ou seja, a especificação da informação para um domínio de aplicação. Os oito esquemas são: semântico, espacial, qualidade, metadados, posicionamento directo, identificadores geográficos, transferência e pesquisa e alteração.

Dos oito esquemas referidos anteriormente, destaca-se a importância do esquema semântico, esquema espacial e esquema de identificadores geográficos neste trabalho. O esquema semântico porque permite modelar a integração da informação geográfica e não geográfica, através da identificação das entidades, atributos e respectivos relacionamentos para um dado domínio de aplicação. O esquema espacial porque permite definir as primitivas geométricas e topológicas das entidades geográficas. As primitivas geométricas que podem ser utilizadas são, por exemplo, o ponto, a linha ou o polígono. Sempre que estas primitivas se encontrem relacionadas umas com as outras, passam a possuir topologia. Para o caso em estudo, e uma vez que o posicionamento é conseguido através da utilização de identificadores geográficos e não de coordenadas, apenas será utilizada a parte do esquema que diz respeito às primitivas topológicas. Tal permitirá conhecer os relacionamentos existentes entre as entidades representadas pelos identificadores utilizados.

O esquema de identificadores geográficos permite explicitar a hierarquia e relacionamentos existentes entre os identificadores utilizados. Tal é conseguido através da implementação de um sistema de identificadores geográficos (Figura 3). No sentido de exemplificar o processo de interpretação das normas e a subsequente incorporação da informação no processo de DCBDG, descreve-se o procedimento utilizado para o esquema de identificadores geográficos, sendo semelhante para os restantes esquemas.

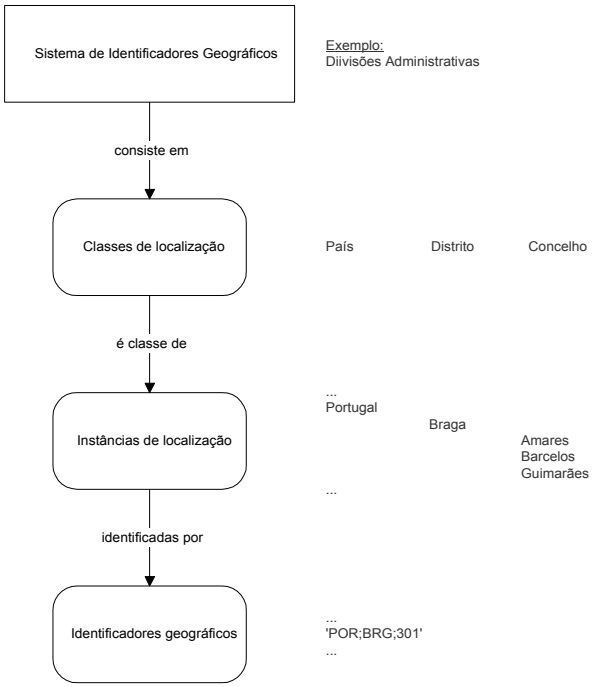

Figura 3 – Sistema de Identificadores Geográficos (Adaptado de CEN/TC-287, *Geographic Information: Referencing, Geographic Identifiers*, Comité Europeu de Normalização, prENV 12661, 1998, p. 6)

A descrição do sistema de identificadores geográficos inclui a descrição do sistema propriamente dito e a descrição de cada *classe* de localização utilizada. A implementação do esquema de identificadores geográficos permite a qualquer utilizador conhecer todos os pormenores associados ao sistema de referenciação em causa. Sendo mandatária a construção do catálogo geográfico para todas as *instâncias* de localização utilizadas, os relacionamentos geográficos existentes entre as mesmas ficam assim devidamente especificados. A interpretação dos diagramas EXPRESS especificados na norma e que definem os requisitos a implementar para o sistema de identificadores geográficos e catálogo geográfico, conduziu à construção do diagrama entidades-relacionamentos apresentado na Figura 4, o qual foi posteriormente implementado numa BD relacional.

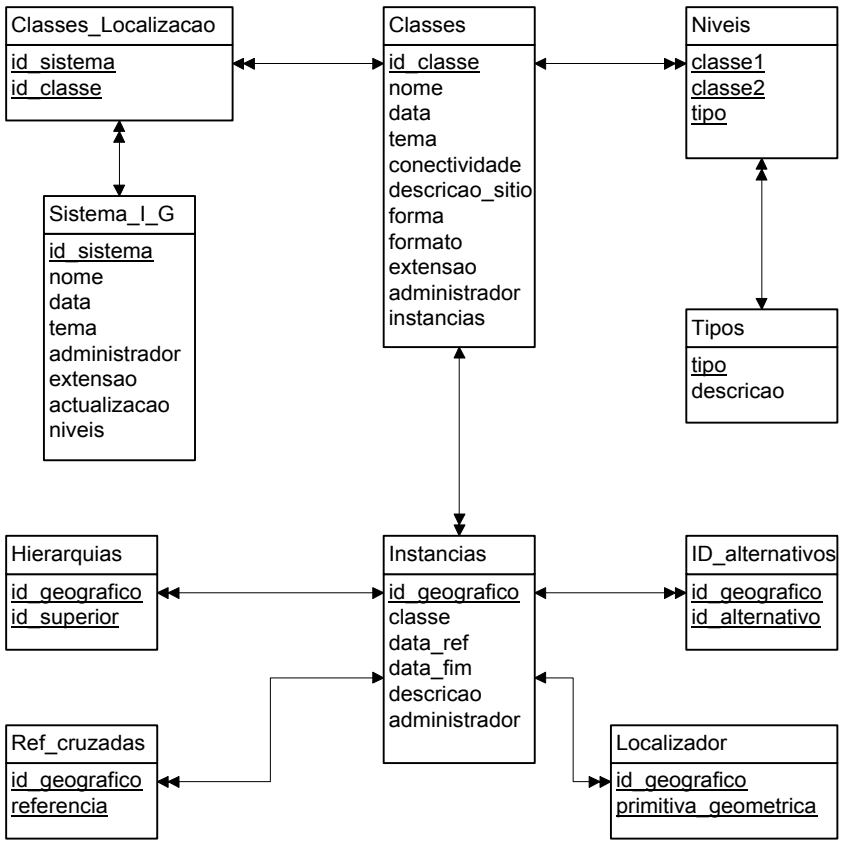

Figura 4 – Esquema de Identificadores Geográficos

O sistema de identificadores geográficos utilizado neste trabalho consiste na adopção de uma divisão administrativa de Portugal que considera subdivisões do território em Concelhos e Distritos (sendo um Distrito uma agregação de vários Concelhos). A implementação deste sistema gera uma BD geográfica onde o conjunto de informação mínimo exigido pela norma CEN TC 287 se encontra especificado. A partir desta informação básica é possível inferir informação geográfica desconhecida para o sistema. No sentido de exemplificar o tipo de informação armazenada no sistema, a Tabela 1 apresenta pequenos extractos de algumas das tabelas/atributos que constituem o diagrama apresentado na Figura 4.

O processo de descoberta de conhecimento inicia-se com uma fase de aprendizagem que é precedida da fase de selecção dos dados, na qual atributos com carácter meramente informativo, e como tal sem qualquer valor na fase de *Data Mining*, deverão ser excluídos. Seguindo este princípio algumas das tabelas (e/ou atributos) apresentadas na Figura 4 não são consideradas nas fases seguintes.

Instancias

id_geografico	Classe
…	…
EVR	Distrito
AVR	Distrito
BRG	Distrito
101	Concelho
104	Concelho
301	Concelho
…	…

Hierarquias

id_geografico	id_superior
…	…
101	AVR
104	AVR
201	BJA
301	BRG
303	BRG
701	EVR
…	…

ID_alternativos

id_geografico	id_alternativo
…	…
AVR	Aveiro
BRG	Braga
PRT	Porto
107	Espinho
201	Aljustrel
301	Amares
…	…

Tabela 1 a) extracto de *Instancias* b) extracto de *Hierarquias* c) extracto de *ID_alternativos*

A aplicação de DCBD utilizada neste trabalho é o *Clementine* da *Integral Solutions, Limited* [ISL 1998a; ISL 1998b]. Esta aplicação permite utilizar dados armazenados em qualquer SGDB, desde que o mesmo suporte uma *Open Database Conectivity* (ODBC). O *Clementine* disponibiliza diversos algoritmos de *Data Mining* (classificação, regressão, *clustering*, …) implementados por diversas técnicas como seja a utilização de regras e árvores de decisão, redes *Kohonen*, redes neuronais, etc. A ferramenta possui ainda um ambiente de programação visual, no qual a construção de uma *stream* que esquematize o processo de DCBD, desde a selecção dos dados até a visualização de resultados, é facilmente conseguida.

Com o objectivo de exemplificar o tipo de resultados que é possível obter utilizando esta abordagem, o *Clementine* foi utilizado na construção de uma *stream* que permite inferir a direcção entre Distritos a partir da direcção existente entre alguns dos Concelhos que constituem a BD geográfica (salienta-se que para os 275 Concelhos de Portugal Continental, agrupados em 18 Distritos, apenas 947 relações se encontram explicitas na BD, o que quer dizer que a maioria dos relacionamentos existentes entre Concelhos são desconhecidos para o sistema). A Figura 5 apresenta a *stream* construída e os resultados obtidos com a utilização do algoritmo C5.0 (refere-se que este algoritmo permite a construção de árvores de decisão, nas quais é inferido o valor do *atributo de saída* baseado nos casos existentes para os *atributos de entrada*).

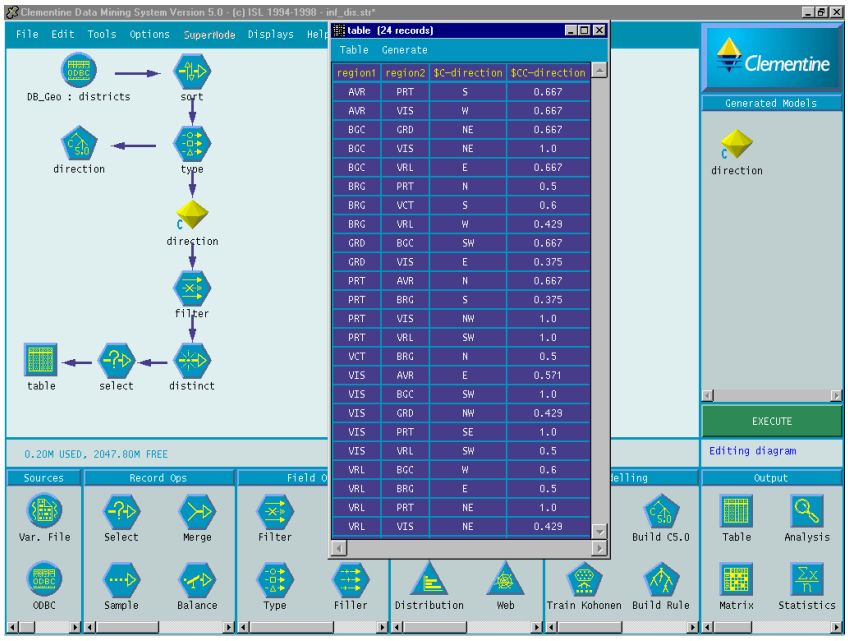

Figura 5 – Inferência de informação geográfica no *Clementine*

Pela análise da Figura 5 é possível verificar que $C-direction representa o atributo inferido pelo modelo criado (neste caso a direcção entre dois Distritos) e $CC-direction evidencia a confiança do resultado. Apesar de em alguns casos a confiança do resultado ser inferior a 0.5, todos os relacionamentos inferidos estão correctos (este valor provém do facto do algoritmo ter de inferir novos relacionamentos baseado num reduzido número de casos disponíveis para o nível hierárquico inferior – o dos Concelhos). O relacionamento inferido para alguns dos Distritos considerados pode ser verificado graficamente pela análise da Figura 6.

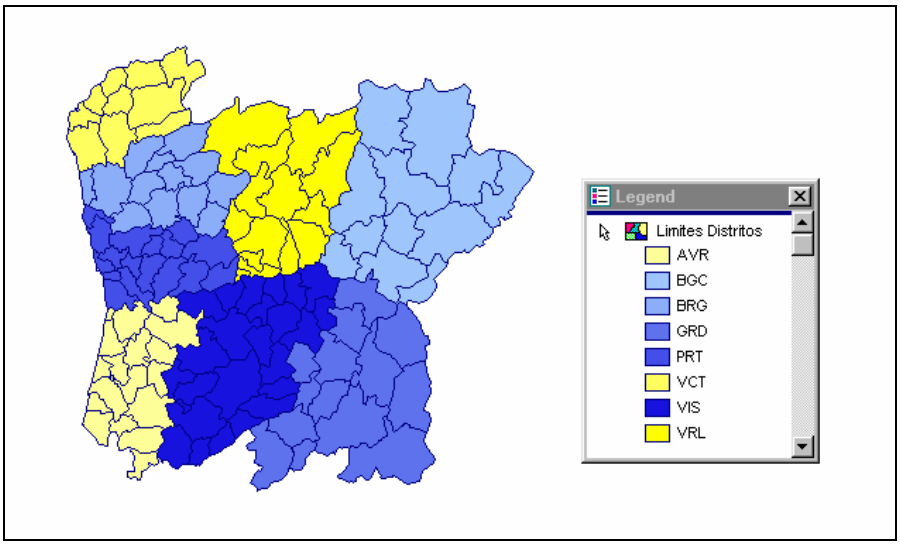

Figura 6 – Localização espacial dos Distritos

Resumidamente, salienta-se que a abordagem apresentada neste trabalho e que passa pela utilização de dados geográficos no processo de DCBD através da explicitação semântica apresenta como vantagens:

- A utilização de ferramentas de DCBD em ampla expansão no mercado, permitindo aproveitar todo o *know how* já adquirido no seu desenvolvimento e na sua respectiva utilização;
- Reutilizar o trabalho que irá ser despendido pelas organizações na adopção da norma CEN TC 287. Estando a Sociedade da Informação em ampla expansão, prevê-se que cada vez mais organizações adoptem estas normas com o objectivo de poderem usufruir da informação presente em Serviços de Informação Geográfica e ainda, o poderem também disponibilizar informação geográfica nestes serviços;
- Considerar a análise espacial no processo de tomada de decisão, através do aproveitamento das características geográficas de alguns atributos presentes na maioria das bases de dados organizacionais (morada, código postal, …).

6 CONCLUSÕES E TRABALHO FUTURO

Na última década verificou-se um acentuado desenvolvimento das capacidades de geração e armazenamento de dados. A possibilidade de extrair, a partir desses dados, conhecimento implícito relacionado com a organização e que pode ser utilizado quer no dia a dia das empresas quer na gestão e decisão estratégica da mesma, tem vindo a atrair a

atenção dos investigadores e deu origem a um novo campo de investigação denominado por Descoberta de Conhecimento a partir de Bases de Dados.

Neste trabalho aborda-se a problemática da inclusão de dados geográficos no processo de descoberta de conhecimento através da explicitação semântica dos relacionamentos espaciais existentes entre algumas das entidades analisadas. Para auxiliar o processo de explicitação e torná-lo ao mesmo tempo o mais genérico possível, recorreu-se à utilização das normas CEN TC 287 para Informação Geográfica, nomeadamente no que diz respeito à construção do esquema semântico, esquema espacial e esquema de identificadores geográficos. Estes esquemas são posteriormente integrados, dando origem a um esquema de aplicação que modela e explicita toda a informação respeitante às entidades, atributos e relacionamentos que caracterizam um dado domínio de aplicação.

Com a construção dos referidos esquemas e a sua posterior implementação numa BD relacional, foi possível utilizar o *Clementine*, uma ferramenta de descoberta de conhecimento para análise exploratória de dados relacionais, na inferência de relacionamentos espaciais implícitos na base de dados geográfica utilizada, a partir de um número reduzido de relacionamentos espaciais explícitos. Esta abordagem além de evitar a necessidade de explicitação de todos os relacionamentos espaciais possíveis entre todas as entidades consideradas no sistema, permite às organizações utilizarem ferramentas que se encontram amplamente divulgadas o mercado (a maioria dos algoritmos de *Data Mining* desenvolvidos para analisar dados geográficos ainda se encontram na fase de *research prototype*).

A experiência adquirida nesta primeira fase de um projecto que se pretende seja cada vez mais abrangente, permite definir como próxima etapa o alargamento do esquema de aplicação agora desenvolvido e que considera apenas um tipo de relacionamento espacial, a direcção, por forma a integrar outros relacionamentos como seja a distância e topologia entre objectos espaciais.

7 REFERÊNCIAS

Abraham, T., e J. F. Roddick, *Discovering Meta-Rules in Mining Temporal and Spatio-Temporal Data*, Proceedings of the 8th. International Database Workshop, Hong Kong, 1997.

Abraham, T., e J. F. Roddick, "Opportunities for Knowledge Discovery in Spatio-Temporal Information Systems", *Australian Journal of Information Systems*, 5, 2 (1998), 3-12.

Andrienko, G. L., e N. V. Andrienko, *Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment*, Varenius Workshop on Status and Trends in Spatial Analysis, Sta. Barbara, CA, 1998.

CEN/TC-287, *Geographic Information: Data Description, Spatial Schema*, Comité Europeu de Normalização, prENV 12160, 1996a.

CEN/TC-287, *Geographic Information: Reference Model*, Comité Europeu de Normalização, prENV 12009, 1996b.

CEN/TC-287, *Geographic Information: Data Description, Metadata*, Comité Europeu de Normalização, prENV 12657, 1998a.

CEN/TC-287, *Geographic Information: Data Description, Quality*, Comité Europeu de Normalização, prENV 12656, 1998b.

CEN/TC-287, *Geographic Information: Data Description, Transfer*, Comité Europeu de Normalização, prENV 12658, 1998c.

CEN/TC-287, *Geographic Information: Data Descrition, Rules for application schemas*, Comité Europeu de Normalização, WI 006, 1998d.

CEN/TC-287, *Geographic Information: Fundamentals, Overview*, Comité Europeu de Normalização, CR 287002, 1998e.

CEN/TC-287, *Geographic Information: Processing, Query and Update: spatial aspects*, Comité Europeu de Normalização, prENV 12660, 1998f.

CEN/TC-287, *Geographic Information: Referencing, Direct Position*, Comité Europeu de Normalização, prENV 12762, 1998g.

CEN/TC-287, *Geographic Information: Referencing, Geographic Identifiers*, Comité Europeu de Normalização, prENV 12661, 1998h.

Dey, S., e S. A. Roberts, *Combining Spatial and Relational Databases for Knowledge Discovery*, Proceedings of the first International Conference on Geo-Computation, Leeds, 1996.

Ester, M., A. Frommelt, H.-P. Kriegel, e J. Sander, *Algorithms for Characterization and Trend Detection in Spatial Databases*, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, 1998a.

Ester, M., H.-P. Kriegel, e J. Sander, *Spatial Data Mining: A Database Approach*, Proceedings of the 5th International Symposium on Large Spatial Databases. Lectures Notes in Computer Science, Berlin, Germany, Springer-Verlag, 1997.

Ester, M., H.-P. Kriegel, J. Sander, e X. Xu, "Clustering for Mining in Large Spatial Databases", *KI-Journal, Special Issue on Data Mining*, 1, (1998b),

Ester, M., H.-P. Kriegel, e X. Xu, *A Database Interface for Clustering in Large Spatial Databases*, Proceedings of the first International Conference on Knowledge Discovery & Data Mining, Montréal, AAAI Press, 1995.

Fayyad, U., G. Piatetsky-Shapiro, e P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, 39, 11 (1996a), 27-34.

Fayyad, U. M., G. Piatetsky-Shapiro, e P. Smyth, *From Data Mining to Knowledge Discovery: An Overview*, in Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, The MIT Press, Massachusetts, 1996b.

Holsheimer, M., e A. Siebes, *Data Mining: The search for Knowledge in Databases*, Technical report CS-R9406, Centrum voor Wiskunde en Informatica - Amsterdam, CS-R9406, 1994.

ISL, *Clementine, Reference Manual, Version 5.0*, 1998a.

ISL, *Clementine, User Guide, Version 5.0*, Integral Solutions Limited, 1998b.

Koperski, K., e J. Han, *Discovery of Spatial Association Rules in Geographic Information Systems*, Proc. 4th International Symposium on Large Spatial Databases (SSD95), Maine, 1995.

Koperski, K., e J. Han, *Data Mining Methods for the analysis of Large Geographic Databases*, Proceedings of the 10th Annual Conference on GIS, Vancouver, 1996.

Koperski, K., J. Han, e N. Stefanovic, *An Efficient Two-Step Method for Classification of Spatial Data*, Proceedings International Symposium on Spatial Data Handling (SDH'98), Vancouver, Canada, 1998.

Lu, W., J. Han, e B. C. Ooi, *Discovery of General Knowledge in Large Spatial Databases*, Proc. of the 1993 Far East Workshop on Geographic Information Systems, Singapura, 1993.

Matheus, C. J., P. K. Chan, e G. Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases", *IEEE Transactions on Knowledge and Data Engineering*, 5, 6 (1993), 903-913.

Schenck, D., e P. Wilson, *Information Modeling: The EXPRESS Way*, Oxford University Press, Oxford, 1994.